Episode Book

EPISODE #6377

How Audio Models Tell Voices Apart

PUBLISHED	31 October 2026 at 14:55 UTC
DURATION	28:23
LISTEN	myweirdprompts.com/episode/speaker-diarization-audio-clustering-42538/
HOSTS	Corn, Herman, Larry, Hilbert
SCRIPT MODEL	openai/gpt-6.1-sol
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.26
SOURCES	17 cited by the research (listed at the end)
BOOK GENERATED	1 November 2026 at 07:41 UTC

SYNOPSIS

Voice recognition systems often struggle with speaker diarization: tracking who spoke when across overlapping speech, whispers, and brief interruptions. This exploration unpacks the mechanics behind pyannote and end-to-end neural diarization (EEND), examining how voice activity detection, speaker embeddings, powerset segmentation, and Bayesian hidden Markov models like VBx separate anonymous speakers. It also looks at the trade-offs of setting fixed speaker counts versus using encoder-decoder attractors to handle unknown voices across challenging recordings.

TRANSCRIPT



Corn

Forty-six point eight percent diarization error on one challenging benchmark. The software can recognise speech and still make a spectacular mess of who owns it. Daniel wants to get underneath that problem, starting with a recording of himself and Hannah. Before attaching names, how does a machine establish that one voice belongs in Speaker One and the other in Speaker Two?



Herman

Especially when either person can return after a silence, interrupt, whisper, or say something as unhelpfully brief as yes.



Corn

A complete contribution, in my view. Daniel's asking about the foundations beneath pyannote, the anonymous separation before anyone is identified. Then he wants to go up a layer, to labels such as Male One and Female One. What acoustic evidence produces those labels, and how much confidence do they deserve?



Herman

Those layers use related signals, but they answer different questions. Anonymous diarization assigns speech to recurring voices. Attribute classification interprets something about those voices. Neither requires a name, and keeping names out lets us see what each system is actually doing.



Corn

Speaker Two has no biography. Refreshing.



Herman

Pyannote is a sensible starting point because it exposes the building blocks, rather than hiding the entire job behind a transcription box. Its own description names speech activity detection, speaker change detection, overlapping speech detection, and speaker embeddings.



Corn

That makes the usual question, which algorithm does diarization, slightly misleading. You're asking which part notices speech, which part notices a change, and which part realises the person coming back is someone we've already heard.



Herman

Diarization partitions speech into stretches associated with a speaker. Segmentation supplies the boundaries and activity. Clustering joins stretches that belong together. The familiar conceptual stack is voice activity detection, change and overlap detection, fixed-dimensional speaker embeddings, then clustering.



Corn

Conceptual matters there. Those needn't be four independent programs passing little envelopes.



Herman

Some functions share a model, and modern systems combine things differently. The community-one pipeline identifies a segmentation model using the powerset formulation, a WeSpeaker embedding model, and VBx clustering. That's a concrete answer to what lives beneath the wrapper.



Corn

Where does actual transcription enter? My guess is it doesn't need to. A machine could tell us we've alternated for ten minutes without understanding a word.



Herman

It can. Words aren't required for the basic task. You can diarize speech in a language the transcriber handles badly. Of course, the acoustic models still have training histories and weaknesses. Language independence doesn't mean immunity to unfamiliar pronunciation or recording conditions.



Corn

Herman, would the machine even know which cluster to call Speaker One?



Herman

Only by some convention, usually the order in which speakers appear. The number is bookkeeping. Swap the labels consistently and the diarization is equally correct.



Corn

An identity crisis solved by renumbering.



Herman

The embedding is the useful unit to hold in mind. A neural network takes a stretch of speech and produces a fixed-length vector. Training encourages recordings of the same speaker to occupy nearby regions, while different speakers should separate. Clustering then works on those vectors, rather than the original waveform.



Corn

Should separate. If Hannah changes from talking normally to laughing, I'd expect a fairly substantial movement in that space. Meanwhile another person with similar vocal anatomy might land inconveniently close.



Herman

Both happen. The network is trying to retain speaker information while discounting phonetic content, room acoustics, microphones, emotion, and so on. It never removes all of that. Short speech is especially awkward because a tiny fragment gives it less evidence about the speaker.



Corn

One syllable and an identity assessment. Efficient hiring practice.



Herman

Would you rather the system leave a short yes unassigned, Corn, or force it into one of the existing clusters?



Corn

Leave it unassigned if the alternative is pretending. But I'd want to distinguish unknown from not enough audio. Otherwise the transcript fills up with mysterious visitors who only say mm.



Herman

That distinction changes the interface. A new cluster suggests a new person. An uncertain assignment says the evidence doesn't support choosing among existing people. Many outputs flatten both situations into a confident speaker label.



Corn

The interface can manufacture certainty even when the model has reservations.



Herman

Locally, before global clustering, a segmentation model looks at short windows and estimates which speakers are active over time. The powerset formulation makes the possible outputs sets of active speakers. Silence is one possibility. One local speaker is another. A pair speaking simultaneously is another.



Corn

Not merely pick the loudest person and call that the conversation.



Herman

It can represent overlap explicitly. A transition from one active speaker to another differs from a transition into both being active. That's valuable when Daniel begins speaking before Hannah finishes, because the correct annotation includes two simultaneous voices.



Corn

Those speaker slots are local, though? Otherwise the model has already done the whole job before clustering gets its coat on.



Herman

Local slots, yes. A speaker called the first speaker in one window needn't correspond to that slot in the next. Stitching local evidence into consistent recording-wide identities is a major part of the work.



Corn

There is also a window-length compromise. Too short and you haven't heard enough voice. Too long and you've probably included someone else.



Herman

Segmentation can help select the speech used for an embedding, excluding overlap or other contaminated regions. But yes, the amount and purity of speech matter together. More audio isn't automatically better if half belongs to a different person.



Herman

Now ... the global stage. VBx treats a sequence of embeddings as observations in a Bayesian hidden Markov model. Hidden speaker states explain the observed vectors, and transitions between states describe how the speaker assignment can evolve.



Corn

You have reached hidden Markov models without moving your water glass. That's restraint.



Herman

I'm keeping the derivation out of it.



Corn

The glass appreciates that.



Herman

A useful part of the model is a persistence prior. It favours continuing with a speaker over rapidly switching labels in response to every small acoustic wobble. Evidence can overcome that preference. It just has to be stronger than the fluctuations produced by ordinary speech.



Corn

Someone can change vowels without summoning another guest.



Herman

The persistence prior stops the clustering from flickering frame to frame, and that's really the second half of the whole problem.

L

Larry

A public service announcement. Attention, everyone. Every day, in your own home, two voices speak, and no one, no one, can tell them apart. Confusion spreads. Neighbours are mislabelled. Clustering breaks down. Diarization fails, and with it, civilisation. But there is one responsible choice. VoiceGuard Anonymous Edition, the only speaker embedding so persistent it never flickers frame to frame. VoiceGuard does not care who you are. It simply knows that voice one is not voice two, and it knows it forever. Guaranteed to sort any household into tidy, anonymous speakers with zero names attached and total moral clarity. Ask for VoiceGuard at the counter. It is, frankly, the only grown up option.



Herman

Persistence belongs to the sequence model, not an embedding that somehow knows forever. An embedding is a measurement of a speech segment. Its usefulness depends on that segment, and a transition prior balances acoustic evidence rather than guaranteeing permanent separation.



Corn

Let it go, Herman. The counting problem is still sitting here without a chair.



Herman

Counting is model selection in a clustering system. How many speaker clusters best explain the vectors? Too few and different people merge. Too many and one person gets split into several supposed speakers.



Corn

I'd expect a room change to tempt it into overcounting. The same person starts near the microphone, then speaks from the doorway, and suddenly Speaker Three has entered through acoustics alone.



Herman

That's a plausible cause. A very brief participant can cause the reverse problem, being absorbed into someone else's cluster. Counting depends on separation, duration, recording conditions, and how the algorithm balances competing explanations.



Corn

Does knowing there are two speakers actually help, or merely force the machine to distribute all its mistakes between two buckets?



Herman

It helps when the knowledge is correct. Pyannote lets you specify an exact speaker count, or lower and upper bounds. Otherwise it estimates. Its release discussion explicitly highlights improvements in speaker assignment and counting, while calling segmentation its established strength.



Herman

Would you give it the count in Daniel's recording?



Corn

Yes, if he's sure nobody else speaks. Refusing information you already possess doesn't make the machine more scientific. But I wouldn't silently turn that into a claim that it discovered two people.



Herman

Providing the count removes one uncertainty, not all of them. It can still assign Hannah's speech to Daniel's cluster. Two clusters with the wrong contents remain wrong.



Corn

Two labelled drawers. Everything still in the wrong one.



Herman

There is an alternative family, end-to-end neural diarization, or EEND. Instead of explicit speaker embeddings followed by a separate clustering stage, the network predicts speaker activity streams jointly. Well … it still has internal learned representations. It skips that particular external embedding-and-clustering arrangement.



Corn

Training needs to tolerate swapped output slots, presumably. Otherwise correct separation gets punished because the model called the second person the first.



Herman

That's what the permutation-free objective addresses. Training evaluates the possible correspondences between predicted streams and reference speakers. The arbitrary numbering doesn't become the learning target.



Corn

A sensible refusal to grade the stationery.



Herman

Early EEND work reported roughly twelve percent diarization error on simulated mixtures, versus roughly twenty-nine percent for its conventional clustering baseline. That mattered because overlap was represented in the output, rather than treated as something a single-speaker clustering decision could wish away.



Corn

Simulated mixtures is carrying some weight. You know the source voices, combine them, and control the experiment. A kitchen recording isn't obliged to resemble that construction.



Herman

Those results support the method under that evaluation, not a universal ranking. Its advocates also objected that separate clustering stages weren't trained directly to minimise the final diarization errors. Joint training addresses that mismatch, but doesn't abolish difficult audio.



Corn

Variable speaker count seems awkward for a network producing output streams. How many streams do you give it before hearing the recording?



Herman

Encoder-decoder attractors offer one answer. The model produces a flexible collection of speaker representations called attractors. Each is compared against the sequence's internal representations to produce a speaker activity stream, with an existence decision controlling how many to retain.



Corn

It learns what the group representatives should be, then asks which frames belong with each?



Herman

Broadly, yes. That approach reported about fifteen percent error on CALLHOME with unknown counts, compared with about nineteen percent for an x-vector clustering baseline. It tackles counting inside the learned system rather than accepting a fixed output roster.



Corn

Still an estimate. Just a better-integrated place to make it.



Herman

Newer work keeps combining these families, including VBx inside EEND-vector-clustering systems for many speakers with little speech each. There isn't one canonical primitive underneath pyannote. There are several separable jobs, and several ways to distribute them between learned models and statistical inference.



Corn

Before adding Male One and Female One, I'd separate two questions that get muddled. Did it correctly join all the speech belonging to one voice? Then, what attributes does it infer about that voice? A bad attribute label shouldn't be allowed to split an otherwise coherent speaker.



Herman

Nor should two voices merge just because an attribute classifier gives them the same label. Masculine-sounding speech is shared by many speakers. It provides very little identifying specificity.



Corn

Half the meeting cannot become Male One.



Herman

The strongest acoustic cue for perceived vocal gender is fundamental frequency, F zero, which relates to the rate of vocal-fold vibration. Higher habitual pitch tends to sound more feminine, and lower habitual pitch more masculine.



Corn

Habitual is doing useful work. If you classify me on one alarmed squeak, you have classified the alarm.



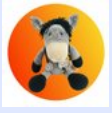
Herman

A stretch of speech gives a distribution of pitch rather than one immutable value. Pitch tracking also has to handle unvoiced sounds and measurement errors. A classifier may learn directly from speech representations instead of receiving a hand-built pitch summary, but those acoustic patterns are still present.



Corn

Herman, your DJ explanation of a voice is usually to turn something down. That won't fix this.



Herman

It fixes more than you'd admit. But changing gain won't change vocal-fold vibration, and even changing pitch doesn't necessarily change all the other cues in the way you expect.



Corn

There's the voice produced at the source, then the resonances imposed by the vocal tract. Formants. I can see how raising pitch alone leaves the other information saying something different.



Herman

Formants are resonant frequency regions shaped by the vocal tract. Their heights and spacing influence perceived gender, with higher formants and wider spacing generally contributing to more feminine perception. They're distinct from the fundamental frequency.



Corn

That's why a pitch-shifted voice can sound altered rather than convincingly like a different kind of speaker. You've moved one set of clues and left others, or shifted everything together unnaturally.



Herman

The details depend on the processing, but that source-and-filter distinction matters. The first several formants contain information about vocal-tract configuration, while pitch comes from the vibrating source. Articulation changes resonances during speech, so again, nobody gets one perfectly stable measurement.



Herman

What would you label if the pitch leaned masculine and the resonances leaned feminine, Corn?



Corn

I'd describe the measurements if that's what the application needs. If it just wants readable speaker labels, I'd stop at Speaker One. Making the display more descriptive isn't automatically making it more accurate.



Herman

You'd also avoid turning disagreement between cues into an invented biological certainty. Testosterone puberty tends to lengthen and thicken the vocal folds, lowering average pitch. But that history includes most cisgender men, some trans women, many trans men, and many non-binary people.



Corn

The machine hears consequences of anatomy and vocal behaviour. It doesn't hear the person's answer to a question about themselves.



Herman

Training and habitual voice use matter too. Pitch, resonance, and delivery can change without moving neatly together. A model trained to map that evidence into two labels will still emit one, even when the person or voice doesn't fit its categories.



Corn

A compulsory choice looks remarkably decisive.



Herman

Breathiness is another cue, though less consistent. Some perceptual results associate it with femininity; others find little significant contribution. Prosody matters as well. Greater pitch variation and dynamic vocal intensity are generally perceived as more feminine in the literature being summarised.



Corn

Those last cues sound especially vulnerable to setting. Someone reading instructions at work and the same person telling a story at dinner might use very different contours.



Herman

They are context-sensitive. Language, speaking style, task, and listener expectations all matter. Rosi and McGettigan's review calls pitch the most robust predictor, rather than claiming every cue is equally strong or universal.



Corn

Robust predictor of perceived gender. Not a microphone-based census.



Herman

Right … perceived vocal gender and self-described gender are different targets. Even within perceived gender, you can ask listeners for a binary choice or ask for ratings of masculinity and femininity. Those produce different statistical problems.



Corn

A benchmark score only makes sense after that choice. Otherwise accurate at what gets lost beneath the percentage.



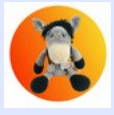
Herman

A WavLM-based classification paper reported over ninety-nine point eight one percent accuracy across its datasets. That's an impressive result for the supplied binary labels. It doesn't establish equivalent performance across every voice, recording environment, or interpretation of gender.



Corn

Ninety-nine point eight one. A very polished decimal.



Herman

The decimal can be legitimate and the extrapolation still wrong. Clean binary-labelled datasets often contain strongly separable vocal populations. A classifier can learn that separation exceptionally well without learning how to represent people outside it.



Corn

We shouldn't sneer at the success, either. For Daniel's stated situation, a clean recording with a male and a female voice, those broad acoustic differences probably provide useful evidence. The mistake is treating that convenient case as the complete population.



Herman

Useful evidence, yes, though that paper doesn't give us a measured accuracy for Daniel and Hannah's recording. Their voices, microphone placement, overlap, and speech duration still determine the actual task.



Corn

What I don't buy is comparing that percentage directly with a correlation score and declaring the model ninety-nine percent certain but only sixty-three percent correct. Different denominators, different targets.



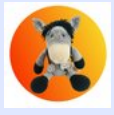
Herman

You shouldn't buy it. A separate perceived-voice study found correlations with human ratings up to zero point six three for female voices and zero point seven seven for male voices. Those are associations between continuous scores, not classification accuracies.



Corn

Hm.



Herman

Squaring the first correlation gives roughly forty percent shared variance in that simple relationship. That leaves substantial variation unaccounted for. But the comparison with binary accuracy is illustrative, not a head-to-head trial of the same system.



Corn

The continuous task exposes differences that the binary task is designed to discard. Two voices can receive the same category while sounding quite different in masculinity or femininity.



Herman

That study's best duration was about seven seconds. Duration is another reminder that we're estimating from a sample, not reading a permanent property off a waveform. More speech can stabilise some measurements, but the particular task and model decide how useful it is.



Corn

Seven seconds. Longer than most people allow a voice before judging it.



Herman

Human judgments are also variable. Training a model to agree with listeners means learning listener perceptions, including their expectations. Agreement with that target doesn't convert perception into someone's identity.



Corn

There's a product-design problem beyond adding uncertainty. Suppose I get ninety-eight percent confidence for Male One. That can mean the model strongly prefers one of its two available bins. It doesn't mean there's only a two percent chance the label misrepresents the person.



Herman

Confidence needs calibration against a defined target and population. If neither available category represents the right answer, stronger confidence merely makes the forced choice firmer. Adding an abstention option can help, but doesn't make identity acoustically observable.



Corn

Male, female, uncertain still leaves uncertain doing several incompatible jobs. Bad microphone. Too little speech. Mixed acoustic cues. A person who doesn't use either identity. The software should not pretend those are one phenomenon.



Herman

The Beyond the Binary review puts the identity boundary plainly: you ask people rather than determine their gender from their voices. Its criticism is about what the measurement can establish. Acoustics can describe and predict perceptions; self-description supplies a different kind of information.



Corn

Does that alter what you'd actually show for Daniel's recording?



Herman

I'd keep anonymous diarization as the primary result. If a separate attribute classifier is useful, I'd label its output as an acoustic inference and keep it separate from speaker identity. If Daniel already knows the participants' descriptions, there's no reason to replace those with a model's guess.



Corn

No prize for rediscovering your wife statistically.



Herman

The literature has another limitation: a heavy focus on transfeminine voices. Even work explicitly about voice diversity doesn't automatically cover all gender-diverse populations evenly. Error analysis on voices that confuse binary classifiers is valuable precisely because the easy majority can conceal systematic mislabelling.



Corn

That's more informative than averaging the failures away. If one kind of voice repeatedly gets the wrong label, another million easy examples don't repair the error for that speaker.



Herman

Anonymous diarization has its own easy-case illusion. Community-one's reported error is forty-four point six percent on AVA-AVD and forty-six point eight percent on Ego4D. Those difficult audio conditions are a long way from a clean, seated conversation.



Corn

We should unpack error before anyone assumes it means the speaker count was wrong that often.



Herman

DER combines missed speech, false alarms, and speaker confusion. Overlapping speech also affects the accounting. It isn't the percentage of files with the wrong head count, and those aggregate numbers don't isolate counting as the sole cause.



Corn

A clean two-person recording helps on several fronts at once. Fewer candidate speakers, more recurring evidence for each, fewer unfamiliar sound conditions. Then we congratulate the whole stack as though we'd tested a crowded doorway.



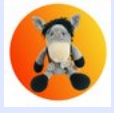
Herman

Even the desired output can simplify reality. Community-one provides regular diarization and an exclusive version intended to reconcile with transcription timestamps. Exclusive output permits only one active speaker at a time.



Corn

The argument becomes orderly in post-production.



Herman

Useful for matching transcript words to speakers, but overlap hasn't disappeared from the recording. Choosing an exclusive representation is choosing which information the downstream application receives.



Hilbert

Zoom H one essential. One hundred and five pounds. Martin at the choir has it, in the cupboard behind the spare music. It's the split-box. I'm paying for it and he's deciding where it lives.



Corn

Possession by convenient cupboard.



Hilbert

He records rehearsals on the split-box. Then he sends the files to a service that labels the speaking between songs. This month it keeps making two of me. One when I'm beside him. Another when I'm by the door telling him the room is still booked for us. He thinks somebody else is helping with the booking.



Herman

The distance change gives it a different recording of your voice. Especially if the doorway changes the reverberation and level.



Hilbert

Corn, I mark every bit of my speech in green pencil on the printout. Even the coughs. Martin says the coughs don't count. They count when the service charges me for another pass. Eighteen pounds so far. I'm keeping that figure at the top of the page.



Corn

Green pencil is becoming a subscription.



Hilbert

There's a separate problem with Ruth. She sings alto, but the talking labels aren't about the singing. The service calls her Male Two when she's reading the room bookings. She crosses it out. Martin runs it again. It comes back. Herman, your distinction between pitch and resonances takes longer to say than their entire label. Hearing you explain the



Hilbert (continued)

doorway changes what I think is happening to my second voice. I'm not asking Martin who else is on it now.



Herman

Running the same classification again doesn't supply information about Ruth that wasn't present the first time.



Hilbert

She's writing her name over it before he scans the sheets. The split-box is still at his place. He's sending another file tonight. The last rehearsal takes twenty-four minutes to correct, and there are six people speaking.



Corn

Twenty-four minutes of correction is the test I'd carry forward. Can the next systems expose an uncertain speaker assignment without forcing someone to rebuild the conversation by hand? Daniel's clean two-person case gives the stack favourable conditions. The useful advance is preserving that usefulness when the evidence gets thin, rather than merely printing firmer labels.



Herman

Thanks to Hilbert Flumingtop for producing the show.



Corn

If you enjoyed this one, try episode seven, Building Custom ASR Tools; episode fifteen, AI Gets Personal; and episode ten, How ASR Went From Frustration To … Whisper Magic. This has been My Weird Prompts, the human-AI collaboration podcast. No question too obscure, no rabbit hole too deep.



Herman

If you enjoy following the machinery underneath the familiar tools, leave us a review wherever you listen. It helps other curious people find the show.



Corn

See you tomorrow.

SOURCES

What the research for this episode read before the script was written. Primary sources first.

1. **pyannote-audio repo, release 4.0.7 (2026-06-30), last push 2026-10-25 [primary]**
   https://github.com/pyannote/pyannote-audio

2. **model card, created 2025-04-15, last modified 2025-09-29 [primary]**
   https://huggingface.co/pyannote/speaker-diarization-community-1

3. **model card, created 2023-11-16, last modified 2024-05-10 [primary]**
   https://huggingface.co/pyannote/speaker-diarization-3.1

4. **pyannote blog on community-1 / pyannote.audio 4.0 [primary]**
   https://www.pyannote.ai/blog/community-1

5. **Wikipedia, last edited 2026-09-19**
   https://en.wikipedia.org/wiki/Speaker_diarisation

6. **Landini et al., VBx, submitted 2020-12-29**
   https://arxiv.org/abs/2012.14952

7. **Fujita et al., EEND permutation-free, INTERSPEECH 2019**
   https://arxiv.org/abs/1909.05952

8. **Fujita et al., EEND multi-label reformulation, 2020-02-24**
   https://arxiv.org/abs/2003.02966

9. **Horiguchi et al., EDA-EEND unknown speaker count, INTERSPEECH 2020**
   https://arxiv.org/abs/2005.09921

10. **Yang & Wang, generic neural clustering, 2022-04-18**
    https://arxiv.org/abs/2204.08164

11. **Pálka et al., VBx for EEND-VC, 2025-10-22**
    https://arxiv.org/abs/2510.19572

12. **Rosi & McGettigan, Voice gender diversity, R. Soc. Open Sci. 12(10):251193, published 2025-10-22**
    https://royalsocietypublishing.org/rsos/article/12/10/251193/236071/

13. **Beyond the binary, review of 107 Interspeech papers**
https://arxiv.org/abs/2409.13335

14. **Yang, Thebaud & Dehak, WavLM demographic prediction, CISS 2025**
https://arxiv.org/abs/2502.12007

15. **Chen et al., voice gender scoring / perceived masculinity-femininity**
https://arxiv.org/abs/2102.07982

16. **Sinha et al., SSL layers for age/gender in children's speech, 2026-06-20**
https://arxiv.org/abs/2606.22177

17. **Moving Towards Non-Binary Gender Identification Via Analysis of System Errors in Binary Gender Classification, ICASSP 2023**
https://doi.org/10.1109/icassp49357.2023.10095997